基于关键帧节点自适应分区与关联的行为识别算法 *

刘锁兰 [1, 2]，田珍珍 [1]，顾嘉晖 [1]，周岳靖 [1]

(1. 常州大学 计算机与人工智能学院, 江苏 常州 213164; 2. 南京理工大学 江苏省社会安全图像与视频理解重点实验室, 南京 210094)

摘　要：基于视频的人体行为识别任务中由于大部分画面并不包含重要的判别信息，这对识别应用的准确性造成严重干扰。关键姿态帧既能表达视频又能降低计算量，且骨骼数据相比于图像包含更多维度的信息。因此，提出一种基于关键帧骨骼节点自适应分区与关联的行为识别算法。首先构建自适应池化深度网络以评估帧的重要性获取关键姿态帧序列；其次通过节点自学习模型建立非自然连接状态下的节点间关联；最后将改进的时空信息应用于 STGCN 并使用 SoftMax 分类识别。在开源的大规模数据集 NTU-RGB+D 和 Kinetics 上与几种典型技术进行比对，验证了所提方法在减少冗余数据量的同时能保留关键动作信息，且动作识别准确率平均提高了 0.63%~11.81%。

关键词：行为识别；关键姿态；自适应；节点关联；STGCN

中图分类号：TP391.41　　doi: 10.19734/j.issn.1001-3695.2022.02.0123

Action recognition based on adaptive partition and association of key-frame nodes

Liu Suolan[1, 2], Tian Zhenzhen[1], Gu Jiahui[1], Zhou Yuejing[1]

(1. *School of Computer& Artificial Intelligence*, *Changzhou University*, *Changzhou Jiangshu* 213164, *China*; 2. *Jiangsu Key Laboratory of Social Security Image & Video Understanding*, *Nanjing University of Science & Technology*, *NanJing* 210094, *China*)

Abstract: In the task of human behavior recognition, most of the video frames do not include important discrimination information, which seriously affects the accuracy of application. Key pose frames can effectively express the video and reduce the amount of computation. Furthermore, bone data contains richer information than RGB image. Therefore, this paper proposes an action recognition approach based on adaptive partition and association of key-frame nodes. Firstly, it constructs an adaptive pooled deep network to evaluate frames importance and obtain key pose sequence. Then, it establishes association between nodes in unnatural connection state by self-learning model. Finally, it applies the improved spatio-temporal information on STGCN and uses SoftMax for classification. This paper evaluates the effectiveness of the proposed approach by comparing with several typical technologies on the open-source and large-scale datasets of NTU-RGB+D and Kinetics. Experimental results show that it can reduce the amount of redundant data and retain key action information, and obtain higher average accuracy by 0.63% ~ 11.81% than the compared methods.

Key words: action recognition; key poses; adaptive; node association; stgcn

0　引言

近些年国内外学者们提出了大量的视频行为识别方法，并给出一些公开的行为数据集[1~5]。图卷积神经网络(graph convolutional network, GCN)起源于卷积神经网络(convolutional neural network, CNN) [6,7]。它将神经网络进行一般化的处理后应用于拓扑图结构中，以代替图的正则化、核方法等传统的拓扑图处理方法，其卓越的性能在视频应用领域得到广泛关注[8~10]。Caramalau 等人[11]将 GCN 应用于人体行为识别任务，通过序列图卷积网络主动学习训练，并应用于训练数据和采样处理，以识别并丢弃多余的未标注数据流得到有效的标注样例。该方法在后续的识别、分类、预测等系列任务中发挥出了比传统方法更好的性能。

相比于其他模态在面对复杂背景、多视角以及遮挡时会出现鲁棒性不足，同时还会产生更多的计算消耗，人体骨骼信息更为清晰直观且不易受到其他外界因素的干扰，具有相对良好的适应性能[12]。一般使用 2D 或者 3D 坐标进行骨骼表示，具有丰富的空间域和时域信息，加强了相邻关节之间的相关性[13]。如 Vemulapall 等人[14]将骨骼表达成一系列的刚体运动，且用一种特殊的欧式群表示刚体间的 3D 几何关系并映射为李群空间中的点，使用 SVM 可以获得较好的分类效果。随着深度学习的成功应用基于深度学习的骨骼建模方法迅速兴起，越来越多的专家学者使用骨骼模态来进行人体检测和行为识别[15~21]。Wang 等人[15] 提出一种基于视频的行为识别模型(temporal segment network，TSN)。该模型可以将稀疏时间采样策略和视频监督相结合，使用整个视频支持有效的学习。Qiu 等人[16]结合 CNN 和 3D 信息提出了一种利用骨骼序列构建关节的三维轨迹进行三维动作识别。该方法首先将每个骨架序列转换为三个片段，每个片段由若干帧组成，然后利用深度神经网络进行时空特征学习。Yan 等人[19]将图卷积网络扩展到时空图模型(spatial temporal graph convolutional networks，STGCN)，设计出用于行为识别的骨骼序列通用表示。STGCN 模型将节点对应于人体的关节，构建多层时空图卷积并让信息沿着空间和时间两个维度进行整合。该

收稿日期：2022-02-16；**修回日期：**2022-05-04　**基金项目：**国家自然科学基金资助项目(61976028)；江苏省社会安全图像与视频理解重点实验室课题(J2021-2)

作者简介：刘锁兰(1980-)，女，江苏泰州人，副教授，硕导，博士，主要研究方向为计算机视觉与人工智能(liusl@cczu.edu.cn)；田珍珍(1997-)，女，河南郑州人，硕士研究生，主要研究方向为模式识别.；顾嘉晖(1996-)，男，硕士，主要研究方向为目标检测与行为识别；周岳靖(1998-)，男，安徽合肥人，硕士研究生，主要研究方向为计算机视觉、图像处理.

方法在测试动作识别数据集上取得了较大的性能提升，其成果引起了广泛关注。在此基础上，文献[20]根据人体关节和骨骼之间的运动相关性，将骨骼数据表示为有向无环图，设计了一种有向图神经网络(directed graph neural networks，DGNN)，用于提取关节、骨骼及其相互关系的信息，并根据提取的特征进行预测。同时，为了更好地适应动作识别任务，在训练过程的基础上对图的拓扑结构进行了自适应性表示，使其性能得到显著改进。Cheng 等人在文献[21]中结合 CNN 中的 shift 结构，将其引入到 GCN，提出了一种改进的移位图进化网络(shift graph convolutional network，Shift-GCN)来克服这两个缺点。Shift-GCN 不使用传统的图卷积操作，而是由新的移位图操作和轻量点卷积组成，其中移位图操作为空间图和时间图提供了灵活的感受野，同时在时序 TCN 上进行 CNN 的 shift 操作，极大地减少了模型参数和计算复杂度。文献[22]描述了一种新的多流注意增强自适应图卷积神经网络(multi-stream adaptive graph convolutional networks，MS-AAGCN)用于骨架的动作识别。图拓扑可以基于端到端的输入数据统一地或单独学习。这种数据驱动的方法增加了模型的灵活性，使其更具通用性，以适应不同的数据样本。此外，通过时空注意力模块进一步增强自适应图卷积层，使模型更加关注重要的关节、帧和特征。在多流框架下，对关节和骨骼的运动信息进行同步建模，提高了识别准确率。Zhao 等人在文献[23]中提出采用基于节点加权贡献的关键帧提取方法，并结合 STGCN 模型进行多特征融合，可以有效提高识别准确率。2021 年 Liu 等[24]针对人体骨架数据提出了一种自适应注意力记忆机制的图卷积网络(adaptive attention memory graph convolutional networks，AAM-GCN)进行动作识别，且在此算法中使用注意机制从骨架序列中提取关键帧，以获取更具鉴别力的时间特征，

综上研究可以发现，在 GCN 模型基础上结合时间和空间维度信息广泛应用于基于骨架的人体行为识别研究[23~26]。同时，由于视频中大部分帧图像都不包含所做的运动信息(静止)，如果把这些也放入网络进行训练，会对训练过程起到反向作用。因此，为排除干扰和信息冗余问题，关键帧提取成为基于视频行为识别的重要预处理环节。

1　基于 STGCN 的人体行为识别

STGCN 是基于图卷积神经网络并加强了时空联系的一类模型，在基于骨架的动作识别中取得了显著的性能。然而 GCN 模型本身仍存在一些问题。比如，在所有模型层和输入数据上对图的拓扑结构进行启发式设置和固定，这可能不适用于 GCN 模型的层次结构和动作识别任务中数据的复杂性与多样性。虽然双流或多流网络进行了空间邻接矩阵的学习，或通过引入增量式自适应模块来增强空间图的表达能力，但其性能仍然受到模型结构本身的限制。此外，GCN 方法通常计算复杂度相当高，一个动作样本的计算复杂度往往超过 15 GFLOPs[27]。尤其随着增量模块、多流融合策略，以及有向无环图网络等的应用，使得计算复杂度急剧增大，而且提取不同特征也需要巨大的计算开销。

通常，在进行视频动作识别时将序列中的所有帧视为同等重要，这使得不能聚焦于最具代表性的帧导致计算量居高不下。大量的研究工作已证明从视频中提取关键帧图像再进行人体行为分析，在降低冗余数据对计算影响的同时，仍能有效使用姿势信息来描述运动信息，表达行为类别。在日常动作中，以“走”为例，此过程包含多个状态，人体在各个时刻也呈现出不同的姿态。在帧序列中目标对象有时是直立的，其他帧中即便出现姿势变化但由于动作的连贯性导致连续几帧对动作识别的贡献存在冗余。因此，如能就帧图像对动作识别的贡献度进行有效衡量，以提取对识别更具信息量的关键姿态帧将有助于降低计算量。依据此，本文设计了一种深度强化子网络来自动学习和获知序列中不同帧的重要性，使重要的帧在分类中起更积极的作用，以降低计算复杂度。

此外，研究表明基于骨骼的行为识别中每个关节点对动作判别并非同等重要。一些行为动作会跟某些关节点构成的集合密切相关，而另一些行为动作则会跟其他一些关节点构成的集合有关。以“打电话”为例，主要与头、肩膀、手肘和手腕这些关节点密切相关，而与腿部关节的关系很小；但是对于“走”、“踢”这类动作的判别就主要通过腿部节点的关联计算才能完成。依据这一点，本文根据序列当前关键帧时空信息和历史信息优化节点分区模型，通过构建节点多级分区建立非自然连接状态下的关联，实现序列中节点关联可以随时间自适应优化，以提升模型鲁棒性并提高识别精度。

2　本文方法

本文采用基于 STGCN 的方法进行人体行为识别。为减少冗余信息对识别效率的影响，在 STGCN 模型基础上提出了视频关键帧提取及骨架节点关联构建方法。首先将视频数据送入深度强化子网络学习和获知序列中不同帧的重要性得到关键信息帧，并通过姿态估计提取骨骼信息。其次，通过节点自学习模型关联序列中不同节点，以衡量节点运动变化对识别的影响。在 STGCN 模块中通过结合关键节点时间和空间信息生成更高层次的特征图，最后通过 SoftMax 分类器进行动作分类识别。

2.1　关键帧判别与提取

视频相比于静态图像来说包含更加丰富信息。但实际一段视频中可能大部分画面并不包含重要的动作判别信息(如静止)，如果把这些帧图像也放入网络训练，则会对训练过程起反作用。因此，如何有效判别冗余信息并提取关键帧是该领域重要的研究内容。但是，现有的关键帧提取算法没有 ground truth，因此需要根据序列间的关联自动生成关键帧。

本文提出的关键姿态帧判别与骨架提取流程如图 1 所示。a)通过采样从原始 RGB 视频序列中抽取初始化的 M 帧图像；b)获取预选帧图像时空特征，并送入多层感知网络模块计算帧间特征差预测其对动作识别的重要性；c)计算预选帧深度特征，在池化环节以当前池化特征和帧间特征差作为输入，这样可以让网络关注之前未关注到的特征进而判断是否为关键帧，得到仅为关键帧的自适应池化向量 F_{pooled}；d)经深度网络输出关键图像帧，采用 Openpose 姿态估计获得关键帧骨架。所提模型旨在利用帧的时空特征，并通过强化学习预测帧间差异来表达帧的重要性，能有效加强帧的判别性，去除冗余信息。

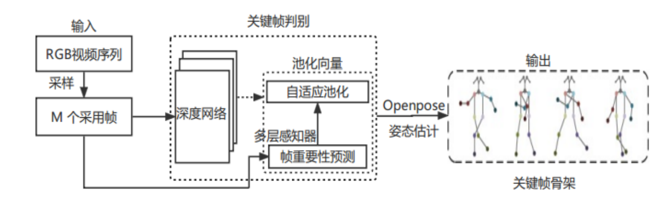


图 1　基于自适应池化网络的关键姿态帧判别模型

Fig. 1　Key pose discrimination model based on adaptive pooling network

设训练样本 $X=\{x_i, y\}$，x_i 为第 i 个训练图像，X 对应的动作类别标签为 y。从训练序列中通过采样得到的初始预选帧集合表示为 $\{a_j\}$，i 与 j 为采样关联。通过深度网络获得帧特征向量为 $\phi(a_j)$。由于采样倾向于“宁多勿少”，因此卷积层输出中仍可能包含大部分冗余信息。在池化环节中引入权重参数，则可有效突出关键帧，同时弱化次要帧的影响。因此通过设计一个具有注意力机制的多层感知网络模型预测每个初

始预选帧的重要性，定义如下：

$$\lambda(a_j)=f_{pre}(\Phi(x_i,t-1),\phi(a_j)) \tag{1}$$

式(1)中 $f_{pre}(\cdot)$ 为采用注意力机制的计算函数， $\Phi(x_i,t-1)$ 表示该帧在原始序列中其前一帧图像的特征，学习过程如下：

$$\Phi(x_i,t)=\lambda(x_i,t-1)\Phi(x_i,t-1)+\lambda(a_j)\phi(a_j) \tag{2}$$

$\lambda(a_i)\in[0,1]$ 表示预选帧重要性权值。

自适应池化过程不仅考虑了预选帧的深度特征，且利用图像帧间信息衡量其所含信息对识别的重要性，得到仅为关键帧的池化向量 F_{pooled} 。经姿态估计处理得到关键帧骨架序列，以便后续行为识别任务。

2.2　节点关联模型

在对视频数据进行关键信息帧姿态估计后构造图 $G=(s,v)$ ，其中 $s\in\mathbb{R}^{N\times3}$ 为包含了 N 个关节点的三维坐标(如 Kinetics 数据集中，采用 Openpose 提取 18 个关节点，则 $N=18$)。人体节点自然关联和固定的分区方案通常不能很好地表达一个序列中发生的所有行为动作。本文设计一种节点关联学习模型，根据序列内容动态优化分区，不仅可以增强/弱化一定邻域范围内节点间的连接，而且可以使没有直接关联的节点之间产生关联。

在 GCN 基础上通过建立节点关联学习模型为节点选择分区，以增强模型的自适应性。

$$f_{out}=W_k f_{in}(S_k+T_k+Q_k) \tag{3}$$

其中 W_k 为权值，S_k 表示基础邻接矩阵即关节点之间的自然连接。单帧骨架图中的关节 s_i 和 s_j 通常存在内在关联和外在关联两种自然依赖关系，可以通过设置不同的参数来进行区分，如式(4)所示。

$$S_k^{ij}=\begin{cases}0, & i=j\\ \vartheta, & connected\\ \rho, & disconnected\end{cases} \tag{4}$$

通过将关节点的自连接设置为 0 可避免自连接对动作判别的影响。内在关联权重用 ϑ 来表示相邻节点之间的物理连接且边距离在动作发生过程中保持不变。外在关联权重用 ρ 表示，指的是节点之间本身不存在物理连接的关系，但在行为过程中却存在较大的联系。例如“打高尔夫球”，左手和右手的自然物理连接并不存在，但是双手共同握住球杆这个关系对于识别此动作具有非常重要的意义。作为可训练的权重邻接矩阵，S_k 亦可通过网络学习数据进行优化，在表达节点之间是否存在联系的同时，对关联的强弱也能进行表达。T_k 表示时间约束，Q_k 表示注意力矩阵。利用注意力机制衡量当前状态图中一节点与其他节点的实际依赖关系，即判断是否连接和连接的强度。因此算法实现的关键是对节点对 (s_i,s_j) 判别动作的实际传入和传出得到依赖权值，计算如下：

$$\xi(s_i,s_j)=\frac{e^{\theta(s_i)^T\gamma(s_j)}}{\sum_{i=1}^{N}\sum_{j=1}^{N}e^{\theta(s_i)^T\gamma(s_j)}} \tag{5}$$

$\theta(\cdot)$ 和 $\gamma(\cdot)$ 分别计算当前节点相对于参考源点的角度和位置信息。为减少参数量、降低计算复杂度，以及防止过拟合，可以通过利用节点的轨迹对任意两个节点间的注意力值进一步计算，并将时间维度融合到 Q_k 中。在训练过程中通常要根据动作类别标签预定义多级分区与关联。如提取 18 个关节点，则关联最多可达 7 级。图 2 所示为以肘部节点为例的 2 级分区和关联示意图。

3　实验与分析

3.1　实验数据集与设置

1) NTU-RGB+D 数据集[4]。该数据集共采集了 56880 个视频样本,包含 drink water、throw、clapping、phone call、shaking hands 等在内的 40 类日常行为动作，9 类与健康相关的动作，以及 11 类双人互动动作。数据采用三个不同水平角度(-45°、0°和 45°)放置的微软 Kinect v2 传感器采集 40 个年龄 10 至 35 岁的人员得到。每个动作执行人做两遍相同的动作。数据形式包括深度信息、3D 骨骼信息、RGB 帧以及红外序列。



（a）骨架图　（b）人体自然结构分区示意（黑色虚线）　（c）2 级分区（红色虚线）及关联示意图（蓝色虚线）

图 2　节点 2 级分区和关联示意图

Fig. 2　Node level-2 zoning and association diagram

数据集提供了两种不分划分标准。a)Cross-Subject 将 ID 为 1, 2, 4, 5, 8, 9, 13, 14, 15,16, 17, 18, 19, 25, 27, 28, 31, 34, 35, 38 共 20 个采集人员的 40320 个视频作为训练集，其余为测试集共 16560 个样本；b)Cross-View 按相机编号划分训练集和测试集。相机 1 采集的 18960 个样本作为测试集，相机 2 和 3 采集的 37920 样本作为训练集。

2)Kinetics-400 数据集[5]。该数据集采集自 YouTube，约 300000 个视频包含了 abseiling、applauding、feeding fish、opening bottle、playing piano、yoga 等在内的人与物体交互动作，以及人与人的互动动作等。涵盖 400 类动作，每类动作至少有 400 个视频样本，每个视频持续约 10 秒。

为降低视频参数差异性对后续处理的影响，本文将所有视频帧分辨率调整为 340×256，同时将帧率转换为 30 frame/s，示例如图 3 所示。



图 3　NTU-RGB+D 和 Kinetics-400 数据集示例

Fig. 3　Samples from NTU-RGB+D and Kinetics-400 datasets

3.2　结果与分析

3.2.1 关键帧提取实验设置与结果分析

在关键帧判别与提取环节，设置 MLP 为四层感知网络，分别采用双曲正切函数 tanh 和最后一层 sigmod 函数为激活函数。通过在模型中加入非线性函数作为隐藏层可以有效解决梯度消失问题。将池化向量的初始状态设置为与第一帧的特征向量一致，后续以当前帧与其前一帧的特征向量差作为自适应模块的输入。

隐藏层神经元数量直接影响感知网络的性能和关键帧提取效果，过少会导致准确度欠佳，过多导致网络过拟合、收敛不理想等问题。同时考虑到不同数据集行为类别数量和样本间的差异性，本文对隐藏层神经元数目 N_{num} 进行动态估算 $N_{num}=\frac{N_{training}}{\partial*(N_i+N_o)}$ 。其中，N_i 和 N_o 分别表示输入层和输出层神经元数目。$N_{training}$ 为训练样本数。调节参数 ∂ 取值范围为 $1\sim10$ 。

为了验证所提关键帧提取算法的有效性，与常用的两种算法进行对比，包括基于运动分析的光流法和视频聚类算法。其中，光流法每次取局部运动光流量最小值作为所要提取的关键帧，聚类法使用 k-means 取距离聚类中心距离最小者为关键帧。实验随机选取 Kinetics 数据集中的 robot dancing 视频片段为例进行说明，结果如图 4 所示。该视频演示动作持

续约 10 秒，图像序列共包含 301 帧。聚类算法提取第 2，19，34 等共计 76 帧为关键帧，光流法提取第 4，48，79，107 等共 16 帧图像作为关键帧。本文算法分别提取第 18，91，199，263，268 共 5 帧图像为关键帧。对比三种方法可见聚类法提取效果较差，尽管压缩率达到 25.1%但仍存在大量冗余，主要原因在于聚类初始化中心数受人为因素干扰严重，且阈值大小的选择也直接导致选取的关键帧数不稳定。光流法主要通过计算镜头中的运动量来反映视频数据中的静止状态，因此该方法对视频镜头的结构选择依赖性较大。视频中第一个动作重复多次出现，本文算法能有效识别并去除重复和冗余，压缩视频，提取 5 帧为关键帧，但却完整反映了视频中的两个关键动作。姿态估计也能有效检测运动目标的关键关节点。

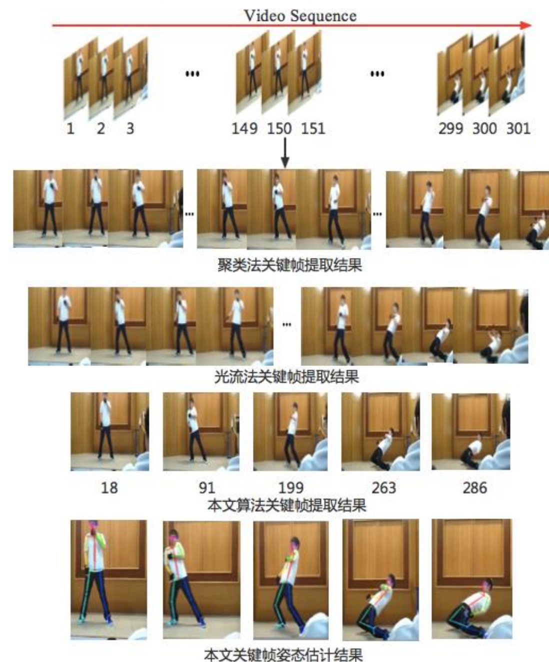


图 4　三种关键帧提取方法结果及本文关键帧姿态估计结果

Fig. 4　Key frame extraction of three algorithms and pose estimation results used in this paper

3.2.2 NTU-RGB+D(Cross-Subject)行为识别结果与分析

将本文工作分别与文献[19]中 STGCN 模型在 uniform，distance 和 spatial 三种分区策略下的识别性能，以及笔者先前报道的研究成果[25]进行了比对。在节点分区和关联算法中设置了 ϑ 和 ρ 两个参数，这里主要通过改变参数值进行实验得到模型的最优性能参数。分别取值 $\vartheta=3$，$\rho=1$ 在加强内在关联的同时适当强化外在关联的影响，同时减少因连接引起的计算量。主要采用 top-1 和 top-5 两个指标对模型性能进行评估。对比方法实验结果为通过设置初始学习率 0.01，在第 80 个 epoch 时减少至初始值的 0.1 倍。本文算法识别率为通过重复实验动态调整每轮迭代相适应的学习率而获得。算法实验环境均为 Ubuntu 16.04 系统，1060-6GB GPU，使用 PyTorch 深度学习框架。

图 5 和 6 分别为在 NTU-RGB+D(Cross-Subject)数据集上的 top-1 和 top-5 的识别率比对。可以看出，模型随着训练逐步优化，本文方法相比于文献[19]和[25]在 top-1 和 top-5 上的识别性能均有一定程度的改进。在第 50 个 epoch 时 top-1 识别精度的最高提升达到 3.84%，在 epoch 为 45 时 top-5 识别精度的最高提升达到 1.34%。在 50 至 80 epoch 区间，识别率基本达到稳定状态分别约为 82%和 97%。这证明对 STGCN 模型改进人体骨骼节点分区和关联可以有效提高模型的识别性能。尤其在对每轮进行相适应的学习率设置之后，本文呈现的实验结果更优于本文之前报道的工作。这也进一步证明了合适的学习率参数对模型性能的改进有着积极作用。

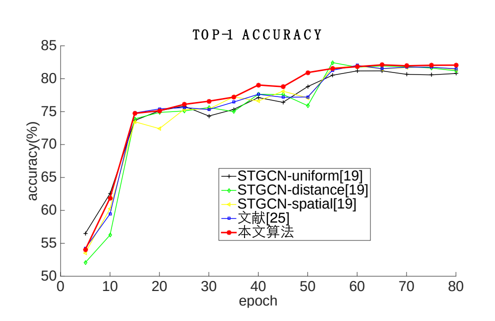


图 5　NTU-RGB+D(Cross-Subject)上 TOP-1 实验结果

Fig. 5　Experimental results of TOP-1 on NTU-RGB+D(Cross-Subject)

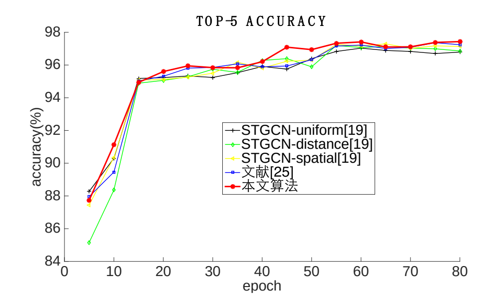


图 6　NTU-RGB+D(Cross-Subject)上 TOP-5 实验结果

Fig. 6　Experimental results of TOP-5 on NTU-RGB+D(Cross-Subject)

3.2.3 NTU-RGB+D(Cross-View) 行为识别结果与分析

表 1 为在 NTU-RGB+D(Cross-View)数据集上的实验比对结果。可以看出本文方法相比于文献[19]的 spatial 方法在其他条件相同下的测试结果在 top-1 和 top-5 上的识别精度分别提升约 2.82%和 0.63%，比文献[25]的最优识别结果提升 5.21%和 1.07%。改进效果明显优于在 NTU-RGB+D(Cross-Subject)上的性能测试，主要原因在于 Cross-View 数据集相对 Cross-Subject 仅改变了数据采集角度，但训练和测试数据来源于全部动作执行人员，极大地降低了动作识别过程中的类内差异。因此，在 NTU-RGB+D(Cross-View)上的识别率与对比方法相当。

表 1　NTU-RGB+D(Cross-View)实验结果对比

Tab. 1　Comparison results on NTU-RGB+D(Cross-View)

epoch	STGCN[19](spatial)		文献[25]		本文算法	
	top-1	top-5	top-1	top-5	top-1	top-5
5	56.96	89.49	59.17	91.04	60.94	92.71
10	65.53	93.15	58.41	89.78	67.02	94.63
15	80.08	97.73	79.55	97.48	76.43	95.77
20	80.68	97.70	81.80	97.84	81.09	96.92
25	83.63	98.27	81.32	97.86	84.37	97.08
30	84.32	98.29	83.02	98.26	87.51	97.82
35	84.95	98.37	82.55	98.08	88.43	98.69
40	**85.90**	**98.54**	**83.51**	**98.10**	**88.72**	**99.17**

3.2.4 在 NTU-RGB+D 数据集上与其他算法的对比

将本文算法模型与当前几种比较典型的方法进行了比较，主要选用了 Lie Group[14]，Deep-LSTM[4]，TSN(Temporal Segment Networks)[15]以及 Clips+CNN+MTLN[16]。其中，Lie Group 方法主要将人体动作表达为流形空间的特征向量。通过建模捕捉帧间时空关联，构成 Lie Group 特征序列形成流形曲线，进行分类识别。Deep-LSTM 网络主要由三层 LSTM 层和全连层(FC Layer)组成。相比于常见的双流网络，TSN 的相主要优势在于解决长时间视频的行为判别，以及小样本导致的过拟合问题。同时该方法对帧序列进行稀疏采样以去除冗余信息降低计算量。Clips+CNN+MTLN 方法首先将骨架序列划分为三个片段，然后使用 CNN 网络学习序列框架中的骨架信息，并使用多任务学习网络(MTLN)联合处理生成的并

行片段，以此合并空间结构信息，识别视频动作。根据数据集划分特点，分别在 Cross-View 和 Cross-Subject 上测试了识别效果。

通过表 2 可以发现本文算法相比于其他几种方法中效果最好的 ClipsCNN+MTLN 在 Cross-View 和 Cross-Subject 的实验方法下的识别精度分别提高了 3.92%和 2.47%。同时，对比几种识别方法可以发现在不同视角下的识别精度整体上均优于不同行为主体实验下的识别精度，最高相差 7.13%。这主要是由于不同执行人即便在采集相同动作时仍因行为习惯等因素导致动作存在较大的差异，这对识别的准确性会产生直接影响。其次，Lie Group 方法更侧重于利用骨骼节点的空间信息而弱化了时序信息对识别的影响，导致效果不理想。Deep-LSTM 方法的优势在于 LSTM 网络对时间序列处理的强大能力，而关键帧的提取则弱化的时间特征，导致识别率不佳。Deep-LSTM，TSN 和 ClipsCNN+MTLN 方法虽然也关注了运动过程中节点的时序信息，但未有效建立节点空间关联且忽略了平移和尺度变化对识别影响。相比于本文方法在压缩视频帧的同时仍通过关键帧保留行为的时序特征，且构建非关联节点在空间的逻辑变化，更能充分表达行为的时空特性，因此在该数据集上表现出优越的识别效果。

表 2 与其他算法的 top-1 最优结果对比

Tab. 2 Best results of top-1 with other compared methods

Methods	X-View	X-Subject
Lie Group[14]	52.80	50.10
Deep LSTM[4]	67.30	60.70
TSN[15]	75.41	78.27
Clips+CNN+MTLN[16]	84.80	79.60
本文算法	**88.72**	**82.07**

3.2.5 Kinetics 行为识别结果与分析

表 3 为将模型改进后在 Kinetics 数据集上的实验结果。可以看出通过设置相适应的迭代学习率之后本文的训练结果在 top-1 和 top-5 上相比于原始 STGCN 模型分别达到了 2.71%和 2.34%的提升。但是对比 NTU-RGB+D 数据集，其测试结果远低于预期，主要原因在于 NTU-RGB+D 采集数据时机位固定且实验环境可控，而 Kinetics 数据集来自于 YouTube 视频采集过程非固定模式，尤其存在大镜头运动导致数据稳定性较差，从而加大了姿态估计难度以及节点间信息关联的难度。但相比于文献[25]所提算法测试结果，可见关键帧的处理以及对人体拓扑结构的关联和加强对此类难度较大的视频数据的识别有着更明显的改善。

表 3 Kinetics 实验结果对比

Tab. 3 Comparison results on Kinetics

epoch	STGCN[19](spatial)		文献[25]		本文算法	
	top-1	top-5	top-1	top-5	top-1	top-5
5	3.56	11.89	5.41	16.29	5.65	17.10
10	4.19	14.23	6.84	20.14	5.96	18.13
15	6.40	18.23	6.59	19.31	14.66	33.62
20	7.11	19.64	7.67	21.63	17.17	37.08
25	14.87	33.88	15.19	34.24	22.77	44.69
30	16.37	35.85	17.26	37.26	23.33	44.66
35	22.77	44.77	22.94	42.31	25.48	47.11

3.2.6 在 Kinetics 数据集上与其他算法的对比

对比方法包括使用特征编码方法(Feature Coding)[28]，以及基于深度学习的 Deep LSTM[4]和 TSN[15]。特征编码方法主要通过从骨架序列中提取人体运动信息，并使用稀疏编码等方法获取特征向量进而行为识别。本节分别比较了几种算法在 top-1 和 top-5 精度方面的识别性能，结果如表 4 所示。所提方法相比于 Deep LSTM 在 top-1 和 top-5 的识别率分别提高了 9.08%和 11.81%。但是在该数据集上的平均识别率整体偏低不到 50%。主要原因在于该视频采集大多基于开放环境且手持设备拍摄不稳定性高，以及大镜头运动造成行为模糊度较高，导致骨架提取难度大。因此，即便构建节点自适应分区与关联模型，识别准确率仍偏低。

表 4 Kinetics 数据集上与其他算法的实验对比

Tab. 4 Comparison results with other three methods on Kinetics

Methods	Top-1	Top-5
Deep LSTM[4]	16.40	35.30
Feature Coding[26]	19.53	36.71
TSN[15]	23.69	44.15
本文算法	**25.48**	**47.11**

4 结束语

在基于视频的行为识别任务中，冗余信息通常会导致模型训练耗时长且对资源需求高，有效识别结果的准确性。实践已证明使用关键姿态帧图像能有效判别行为类别。基于此，本文在 STGCN 模型基础上引入深度强化子网络来学习和评估序列中不同帧的重要性，提取关键帧以表达视频，并通过姿态估计获得骨骼信息。通过节点自适应模型学习非自然连接状态下的节点间关联，以扩展节点的运动变化对识别的影响。在 NTU-RGB+D 和 Kinetics 两个具有挑战性的大规模数据集上的测试效果相比于几种主流的识别技术 Feature Coding、Lie Group、Deep LSTM、TSN 以及 Clips+CNN+MTLN 皆呈现一定程度的提升。值得注意的是节点关联模型虽然能建立非自然分区下的节点联系，但为了降低计算复杂度关联参数 ϑ 和 ρ 的选择不宜过大。本文为在重复实验条件下选择最优参数，因此如能通过建立合适的目标函数进一步自动优化参数选择是后续研究的重要内容。

参考文献：

[1] Zhang Hongbo, Zhang Yixiang, Zhong Bineng, *et al.* A comprehensive survey of vision-based human action recognition methods [J]. Sensors 2019, 19: 1005-1025.

[2] Wang Zhengwei, She Qi, Smolic A. ACTION-NET: Multipath excitation for action recognition [C]. /Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2021. https://doi. org/10. 48550/arXiv. 2103. 07372.

[3] Wang Limin, Tong Zhan, Ji Bin Ji, *et al.* TDN: Temporal Difference Networks for Efficient Action Recognition [C]. /Proc of IEEE Conference on Computer Vision and Pattern Recognition, Piscataway, NJ: IEEE Press, 2021. https://doi. org/10. 48550/arXiv. 2012. 10071.

[4] Shahroudy A, Liu Jun, Tsong N T, *et al.* NTU RGB+D: A large scale dataset for 3D human activity analysis [C]. /Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2016: 1010-1019.

[5] Kay W, Carreira J, Simonyan K, *et al.* The kinetics human action video dataset [EB/OL]. (2017) . https://doi. org/10. 48550/arXiv. 1705. 06950.

[6] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C]. /Proc of International Conference on Neural Information Processing Systems, 2012, 1097-1105.

[7] Ji Shuiwang, Xu Wei, Yang Ming, *et al.* 3D Convolutional neural networks for human action recognition [J]. IEEE Trans on Pattern Analysis and Machine Intelligence (PAMI) , 2012, 35 (1): 221-231.

[8] 徐冰冰，岑科廷，黄俊杰，等. 图卷积神经网络综述 [J]. 计算机学报, 2020, 43 (5): 755-780. (Xu Bingbing, Cen Keting, Huang Junjie, *et al.* A survey on graph convolutional neural network [J]. Chinese Journal of Computers, 2020, 43 (5): 755-780.)

[9] Duvenaud D, Maclaurin D, Aguilera J, *et al.* Convolutional networks on graphs for learning molecular fingerprints [J]. Advances in Neural Information Processing Systems, 2015: 2224-2232.

[10] 谢昭，周义，吴克伟，等．基于时空关注度 LSTM 的行为识别 [J]. 计算机学报, 2021, 44 (2): 261-274. (Xie Zhao, Zhou Yi, Wu Kewei, *et al.* Activity Recognition Based on Spatial-Temporal Attention LSTM [J]. Chinese Journal of Computers, 2021, 44 (2): 261-274.)

[11] Caramalau R, Bhattarai B, Kim T K. Sequential graph convolutional network for active learning [C]. /Proc of IEEE Conference on Computer Vision and Pattern Recognition, Piscataway, NJ: IEEE Press, 2021. https://doi. org/10. 48550/arXiv. 2006. 10219.

[12] Zhao Hang, Torralba A, Torresani L, *et al.* Hacs: Human action clips and segments dataset for recognition and temporal localization [C]. /Proc of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2019: 8668-8678.

[13] Li Kunchang, Li Xianhang, Wang Yali, *et al.* CT-net: Channel tensorization network for video classification [C]. /Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2018. https://doi. org/10. 48550/arXiv. 2106. 01603.

[14] Vemulapalli R, Arrate F, Chellappa R, *et al.* Human action recognition by representing 3D skeletons as points in a lie group [C]. /Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2014: 588-595.

[15] Wang Limin, Xiong Yuanjun, Wang Zhe, *et al.* Temporal Segment networks: Towards good practices for deep action recognition [C]. /Proc of IEEE Conference on European Conference on Computer Vision, 2016. https://doi. org/10. 48550/arXiv. 1608. 00859.

[16] Qiu Hongke, Mohammed B, Senjian A, *et al.* A new representation of skeleton sequences for 3D action recognition [C]. /Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017. https://doi. org/10. 48550/arXiv. 1703. 03492.

[17] Christoph F, Hao Qifan, Jitendra M, *et al.* Slow fast networks for video recognition [C]. /Proc of IEEE International Conference on Computer Vision, 2019: 6201–6210.

[18] Cheng Xiaopeng, Feng Dapeng. Skeleton embedded motion body partition for human action recognition using depth sequences [J]. Signal Processing, 2018, 143: 56-68.

[19] Yan Sijie, Xiong Yuanjun, Lin Dahua. Spatial temporal graph convolutional networks for skeleton-based action recognition [C]. /Proc of the Thirty-second AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2018, 32 (01): 7444-7452.

[20] Shi Lei, Zhang Yifang, Cheng Jian, *et al.* Skeleton-Based Action Recognition with Directed Graph Neural Networks [C]. /Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2019: 7912-7921.

[21] Cheng Ke, Zhang Xifan, He Xiangyu, *et al.* Skeleton-based action recognition with shift graph convolutional network [C]. /Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2020: 183-192.

[22] Shi Lei, Zhang Yifang, Cheng Jian, *et al.* Skeleton-Based Action Recognition with Multi-Stream Adaptive Graph Convolutional Networks [J]. IEEE Trans on Image Processing, 2020, 29: 9532-9545.

[23] Zhao Yuerong, Guo Hongbo, Gao Ling, *et al.* Multifeature fusion action recognition based on key frames [J]. Concurrency & Computation Practice & Experience, 2021, 2021 (3) . https://doi. org/10. 1002/cpe. 6137.

[24] Liu Di, Xu Hui, Wang Jianzhong, *et al.* Adaptive Attention Memory Graph Convolutional Networks for Skeleton-Based Action Recognition [J]. Sensors, 2021, 21: 6761-6780.

[25] 刘锁兰，顾嘉晖，王洪元，等．基于关联分区和 ST-GCN 的人体行为识别 [J]. 计算机工程与应用, 2021, 57 (3): 168-178. (Liu Suolan, Gu Jiahui, Wang Hongyuan, *et al.* Human behavior recognition based on associative partition and ST-GCN [J]. Computer Engineering and Application, 2021, 57 (3): 168-178.)

[26] Jian Meng, Zhang Shuai, Wu Lifang, *et al.* Deep key frame extraction for sport training [J]. Neurocomputing, 2019, 328: 147-156.

[27] 陈煜平，邱卫根．基于视觉的人体行为识别算法研究综述 [J]. 计算机应用研究, 2019, 36 (7): 1927-1934. (Chen Yuping, Qiu Weigen. Survey of human action recognition algorithm based on vision [J]. Application Research of Computers, 2019, 36 (7): 1927-1934.)

[28] Zhang Yixiang, Zhang Hongbo, Du Jixiang, *et al.* RGB+2D skeleton: local hand-crafted and 3D convolution feature coding for action recognition [J]. Signal, Image and Video Processing, 2021, 2021 (15): Article ID 1386.